

Klasterisasi Pelanggan Mall Menggunakan Algoritma K-Means Untuk Mendukung Strategi Segmentasi Pelanggan

Muhamad Ikhsan¹, Wismaningsih²

^{1,2}Program Studi : Sistem Informasi, Universitas Darwan Ali

Email : muhamadikhsan.ktw@gmail.com¹, wismaningsih124@gmail.com²

ABSTRACT— Customer segmentation is an important strategy in modern business to understand consumer characteristics and improve marketing effectiveness. Companies need to identify customer groups with similar purchasing behaviors so that marketing strategies can be tailored to each segment. This study aims to classify mall customers using the K-Means clustering algorithm based on customer characteristics, namely annual income and spending score. The dataset used consists of 500 customer records containing CustomerID, Gender, Age, Annual Income, and Spending Score. The research stages include data collection, data preprocessing, exploratory data analysis, feature selection, data standardization using StandardScaler, determining the optimal number of clusters using the Elbow Method, clustering using the K-Means algorithm, and evaluating the clustering results using the Silhouette Score. The results indicate that the optimal number of clusters is five. Each cluster represents customers with distinct characteristics, such as low-income customers with low spending, low-income customers with high spending, medium-income customers, high-income customers with low spending, and high-income customers with high spending. These clustering results provide valuable information for businesses to design more effective marketing strategies, customer loyalty programs, and promotional campaigns. The implementation of the K-Means algorithm demonstrates its capability to classify customer data accurately and efficiently, making it suitable for customer segmentation problems in retail businesses.

Keywords— K-Means Algorithm, Customer Clustering, Customer Segmentation, Annual Income, Spending Score.

ABSTRAK— Segmentasi pelanggan merupakan salah satu strategi penting dalam dunia bisnis modern untuk memahami karakteristik konsumen dan meningkatkan efektivitas pemasaran. Perusahaan perlu mengidentifikasi kelompok pelanggan yang memiliki pola pembelian serupa sehingga strategi pemasaran dapat disesuaikan dengan karakteristik masing-masing kelompok. Penelitian ini bertujuan untuk mengelompokkan pelanggan mall menggunakan algoritma K-Means berdasarkan karakteristik pelanggan, yaitu pendapatan tahunan dan tingkat pengeluaran. Dataset yang digunakan terdiri atas 500 data pelanggan yang memuat atribut CustomerID, Gender, Age, Annual Income, dan Spending Score. Tahapan penelitian meliputi pengumpulan data, preprocessing, eksplorasi data, pemilihan atribut, standarisasi data menggunakan StandardScaler, penentuan jumlah cluster menggunakan Elbow Method, proses clustering menggunakan algoritma K-Means, serta evaluasi hasil menggunakan Silhouette Score. Hasil penelitian menunjukkan bahwa jumlah cluster optimal adalah lima kelompok pelanggan. Setiap cluster memiliki karakteristik yang berbeda, yaitu pelanggan dengan pendapatan rendah dan pengeluaran rendah, pendapatan rendah dan pengeluaran tinggi, pendapatan sedang, pendapatan tinggi dan pengeluaran rendah, serta pendapatan tinggi dan pengeluaran tinggi. Hasil pengelompokan ini dapat dimanfaatkan oleh perusahaan untuk menyusun strategi pemasaran yang lebih efektif, meningkatkan loyalitas pelanggan, dan menentukan target promosi yang sesuai. Implementasi algoritma K-Means menunjukkan kemampuan yang baik dalam mengelompokkan data pelanggan secara cepat dan efisien sehingga dapat digunakan sebagai salah satu metode pendukung pengambilan keputusan dalam segmentasi pelanggan.

Kata kunci— Algoritma K-Means, Klasterisasi Pelanggan, Segmentasi Pelanggan, Pendapatan Tahunan, Spending Score.

I. PENDAHULUAN

Perkembangan teknologi informasi telah mendorong pertumbuhan data dalam jumlah yang sangat besar di berbagai sektor, termasuk sektor perdagangan dan ritel. Setiap aktivitas pelanggan, seperti transaksi pembelian, penggunaan kartu keanggotaan, hingga interaksi melalui layanan digital, menghasilkan data yang dapat dimanfaatkan sebagai sumber informasi bagi perusahaan. Data tersebut tidak hanya berfungsi sebagai arsip transaksi, tetapi juga dapat diolah menjadi informasi yang mendukung pengambilan keputusan secara lebih objektif dan berbasis bukti. Oleh karena itu, pemanfaatan teknik

Data Mining menjadi salah satu pendekatan yang banyak diterapkan untuk menemukan pola, hubungan, maupun karakteristik tersembunyi dari sekumpulan data yang berukuran besar.

Dalam dunia bisnis ritel, persaingan yang semakin ketat menuntut perusahaan untuk memahami perilaku konsumennya secara lebih mendalam. Strategi pemasaran yang diterapkan secara umum kepada seluruh pelanggan sering kali tidak memberikan hasil yang optimal karena setiap pelanggan memiliki karakteristik, kebutuhan, dan kebiasaan berbelanja yang berbeda. Kondisi tersebut mendorong perusahaan untuk menerapkan konsep

customer segmentation, yaitu proses mengelompokkan pelanggan berdasarkan tingkat kemiripan karakteristik tertentu sehingga setiap kelompok dapat diberikan strategi pemasaran yang sesuai [3], [4]. Melalui segmentasi pelanggan, perusahaan dapat meningkatkan efektivitas promosi, memperkuat loyalitas pelanggan, serta mengoptimalkan alokasi sumber daya pemasaran.

Salah satu metode yang banyak digunakan dalam proses segmentasi pelanggan adalah teknik *clustering*. Berbeda dengan metode klasifikasi yang memerlukan data berlabel, *clustering* termasuk ke dalam pendekatan *unsupervised learning* yang mampu mengelompokkan data berdasarkan tingkat kemiripan antarobjek tanpa memerlukan informasi kelas sebelumnya [2], [5]. Teknik ini sangat sesuai digunakan pada data pelanggan karena pada umumnya perusahaan belum memiliki label atau kategori pelanggan secara spesifik. Dengan demikian, proses pengelompokan dilakukan secara otomatis berdasarkan karakteristik yang dimiliki oleh setiap pelanggan.

Di antara berbagai algoritma *clustering*, algoritma K-Means merupakan salah satu metode yang paling banyak digunakan karena memiliki konsep yang sederhana, proses komputasi yang relatif cepat, serta mampu menghasilkan kelompok data yang memiliki tingkat homogenitas tinggi [6], [7]. Algoritma ini bekerja dengan membagi data ke dalam sejumlah kelompok berdasarkan kedekatan jarak terhadap titik pusat (*centroid*) menggunakan perhitungan jarak Euclidean. Selain mudah diimplementasikan, K-Means juga mampu menangani data dalam jumlah besar sehingga banyak diterapkan pada berbagai bidang, seperti segmentasi pelanggan, analisis pemasaran, kesehatan, pendidikan, hingga analisis perilaku pengguna layanan digital [8].

Pada sektor pusat perbelanjaan (*mall*), pelanggan memiliki karakteristik yang sangat beragam. Sebagian pelanggan memiliki pendapatan tinggi dan melakukan pembelian secara rutin, sementara sebagian lainnya memiliki tingkat pendapatan yang relatif rendah tetapi menunjukkan intensitas belanja yang tinggi. Di sisi lain, terdapat pula pelanggan dengan kemampuan ekonomi tinggi yang justru memiliki tingkat pengeluaran rendah karena preferensi maupun pola konsumsi yang berbeda. Perbedaan karakteristik tersebut menyebabkan perusahaan menghadapi tantangan dalam menentukan strategi pemasaran yang mampu menjangkau seluruh pelanggan secara efektif. Tanpa adanya proses segmentasi yang baik, promosi yang diberikan berpotensi tidak tepat sasaran sehingga mengurangi efisiensi biaya pemasaran dan peluang peningkatan penjualan.

Beberapa penelitian sebelumnya menunjukkan bahwa algoritma K-Means mampu memberikan hasil segmentasi pelanggan yang baik pada berbagai jenis data transaksi. Namun, sebagian penelitian hanya berfokus pada proses pembentukan cluster tanpa melakukan evaluasi kualitas hasil pengelompokan menggunakan metrik validasi.

Selain itu, terdapat penelitian yang menggunakan jumlah data relatif terbatas sehingga variasi karakteristik pelanggan belum sepenuhnya tergambarkan. Kondisi tersebut menunjukkan bahwa masih diperlukan penelitian yang mengombinasikan proses *preprocessing* data, penentuan jumlah cluster yang optimal, serta evaluasi kualitas cluster untuk menghasilkan segmentasi pelanggan yang lebih representatif [9]–[11].

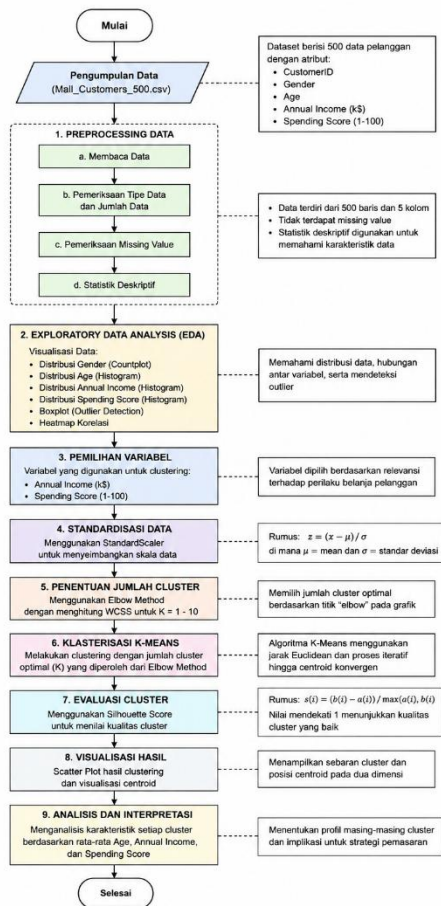
Berdasarkan permasalahan tersebut, penelitian ini menerapkan algoritma K-Means untuk melakukan klasterisasi pelanggan mall menggunakan dataset yang terdiri atas 500 data pelanggan. Variabel yang digunakan dalam proses pembentukan cluster adalah Annual Income (k\$) dan Spending Score (1–100) karena kedua atribut tersebut dianggap mampu merepresentasikan kemampuan ekonomi dan perilaku belanja pelanggan. Sebelum proses klasterisasi dilakukan, data distandardisasi menggunakan metode StandardScaler agar setiap variabel memiliki skala yang seimbang. Selanjutnya, jumlah cluster optimal ditentukan menggunakan Elbow Method, sedangkan kualitas hasil pengelompokan dievaluasi menggunakan Silhouette Score untuk memastikan bahwa cluster yang terbentuk memiliki tingkat homogenitas dan pemisahan yang baik.

II. METODOLOGI PENELITIAN

A. Tahapan Penelitian

Penelitian ini menggunakan pendekatan Data Mining dengan metode Knowledge Discovery in Database (KDD) untuk memperoleh informasi dari data pelanggan mall. Proses penelitian dilakukan secara sistematis mulai dari pengumpulan data hingga evaluasi hasil klasterisasi. Algoritma yang digunakan adalah K-Means Clustering karena mampu mengelompokkan objek berdasarkan tingkat kemiripan karakteristik tanpa memerlukan label kelas sebelumnya.

Tahapan penelitian ditunjukkan pada Gambar 1, yang terdiri atas pengumpulan dataset, preprocessing data, eksplorasi data (*Exploratory Data Analysis*), pemilihan variabel, standardisasi data, penentuan jumlah cluster menggunakan *Elbow Method*, proses klasterisasi menggunakan algoritma K-Means, evaluasi menggunakan *Silhouette Score*, visualisasi hasil, serta analisis karakteristik masing-masing cluster.



Gambar 1. Flowchart Tahapan Penelitian

B. Dataset Penelitian

Dataset yang digunakan merupakan dataset pelanggan mall yang terdiri atas 500 data pelanggan. Setiap baris data merepresentasikan satu pelanggan dengan beberapa atribut yang menggambarkan karakteristik demografis dan perilaku belanja.

TABEL I
DESKRIPSI VARIABEL PENELITIAN

Variabel	Tipe Data	Keterangan
CustomerID	Integer	Nomor identitas pelanggan
Gender	Kategori	Jenis kelamin pelanggan
Age	Integer	Umur pelanggan (tahun)
Annual Income (k\$)	Integer	Pendapatan tahunan pelanggan
Spending Score (1-100)	Integer	Skor tingkat pengeluaran pelanggan

Pada penelitian ini, atribut Annual Income (k\$) dan Spending Score (1-100) digunakan sebagai variabel utama dalam proses klasterisasi karena kedua variabel tersebut dinilai mampu merepresentasikan kondisi ekonomi dan perilaku konsumsi pelanggan. Sementara itu, atribut Age digunakan sebagai informasi pendukung dalam analisis karakteristik setiap cluster, sedangkan

CustomerID hanya berfungsi sebagai identitas data dan tidak dilibatkan dalam proses pembentukan cluster.

C. Preprocessing Data

Tahap preprocessing dilakukan untuk memastikan bahwa dataset memiliki kualitas yang baik sebelum diproses menggunakan algoritma K-Means. Data yang berkualitas akan meningkatkan akurasi hasil klasterisasi dan meminimalkan kesalahan analisis. Proses preprocessing meliputi beberapa tahapan sebagai berikut.

1. Membaca dataset menggunakan pustaka Pandas.
2. Memeriksa jumlah data dan tipe data setiap atribut.
3. Mengidentifikasi keberadaan *missing value*.
4. Menghasilkan statistik deskriptif sebagai gambaran umum data.
5. Memilih atribut yang digunakan pada proses klasterisasi.

Hasil pemeriksaan menunjukkan bahwa dataset tidak memiliki nilai kosong (*missing value*), sehingga seluruh data dapat digunakan tanpa perlu melakukan proses imputasi maupun pembersihan data tambahan.

D. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) dilakukan untuk memperoleh gambaran awal mengenai karakteristik dataset sebelum proses klasterisasi dilakukan. Analisis ini bertujuan untuk mengetahui distribusi data, hubungan antarvariabel, serta mendeteksi kemungkinan adanya data pencilan (*outlier*).

Visualisasi yang digunakan dalam penelitian ini meliputi:

1. Countplot untuk melihat distribusi pelanggan berdasarkan jenis kelamin.
2. Histogram untuk mengetahui distribusi umur pelanggan.
3. Histogram Annual Income untuk melihat sebaran pendapatan tahunan.
4. Histogram Spending Score untuk melihat distribusi tingkat pengeluaran pelanggan.
5. Boxplot untuk mendeteksi data pencilan pada variabel numerik.
6. Heatmap Korelasi untuk mengetahui hubungan antarvariabel numerik.

Tahapan ini membantu peneliti memahami karakteristik data sebelum dilakukan proses pembentukan cluster.

E. Standardisasi Data

Algoritma K-Means menggunakan perhitungan jarak Euclidean dalam menentukan kedekatan antarobjek. Oleh karena itu, variabel yang digunakan perlu memiliki skala yang seimbang agar tidak terjadi dominasi salah satu variabel terhadap variabel lainnya.

F. Penentuan Jumlah Cluster Menggunakan Elbow Method

Algoritma K-Means mengharuskan pengguna menentukan jumlah cluster (KKK) sebelum proses klasterisasi dilakukan. Untuk memperoleh jumlah cluster yang optimal digunakan metode Elbow Method.

Metode ini menghitung nilai Within Cluster Sum of Squares (WCSS) untuk beberapa jumlah cluster. Nilai WCSS dihitung menggunakan Persamaan (2).

G. Algoritma K-Means

K-Means merupakan algoritma *unsupervised learning* yang mengelompokkan data berdasarkan kedekatan terhadap titik pusat cluster (*centroid*). Proses algoritma dilakukan secara iteratif hingga posisi centroid tidak mengalami perubahan yang signifikan.

Jarak antara data dengan centroid dihitung menggunakan Euclidean Distance sebagaimana ditunjukkan pada Persamaan (3).

H. Evaluasi Menggunakan Silhouette Score

Setelah proses klasterisasi selesai dilakukan, kualitas hasil cluster dievaluasi menggunakan **Silhouette Score**. Metrik ini digunakan untuk mengukur tingkat kemiripan suatu data terhadap cluster tempat data tersebut berada dibandingkan dengan cluster lainnya.

Perhitungan Silhouette Score ditunjukkan pada Persamaan (4).

I. Perangkat Penelitian

Penelitian ini dilaksanakan menggunakan perangkat lunak dan perangkat keras yang mendukung proses pengolahan data.

Tabel II
Perangkat Penelitian

Komponen	Keterangan
Sistem Operasi	Windows 10/11
Platform	Google Colaboratory
Bahasa Pemrograman	Python 3
Library	Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn
Metode	K-Means Clustering
Evaluasi	Elbow Method dan Silhouette Score

III. DESAIN, HASIL DAN PEMBAHASAN

A. Hasil Preprocessing Data

Tahap awal penelitian diawali dengan proses *preprocessing* untuk memastikan bahwa dataset berada dalam kondisi yang layak digunakan pada proses klasterisasi. Dataset yang digunakan terdiri atas 500 data pelanggan mall dengan lima atribut, yaitu *CustomerID*, *Gender*, *Age*, *Annual Income (k\$)*, dan *Spending Score (1–100)*. Pemeriksaan dilakukan terhadap jumlah data, tipe data setiap atribut, keberadaan *missing value*, serta statistik deskriptif.

Hasil pemeriksaan menunjukkan bahwa seluruh atribut telah memiliki tipe data yang sesuai. Atribut *CustomerID*, *Age*, *Annual Income (k\$)*, dan *Spending Score (1–100)* bertipe numerik (*integer*), sedangkan atribut *Gender* bertipe kategorik (*object*). Selain itu, hasil pemeriksaan *missing value* menunjukkan bahwa tidak

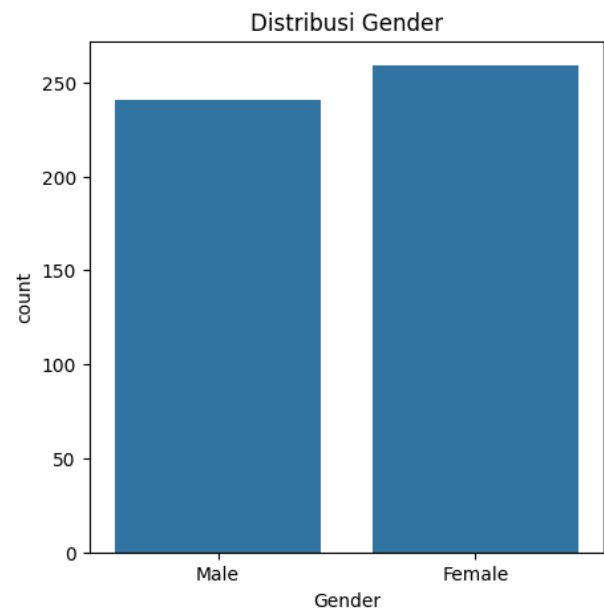
terdapat data yang kosong sehingga seluruh data dapat digunakan pada proses analisis tanpa memerlukan proses imputasi.

Statistik deskriptif juga menunjukkan bahwa distribusi nilai pada setiap atribut masih berada pada rentang yang wajar dan tidak ditemukan nilai yang menyimpang secara ekstrem. Dengan demikian, dataset dinilai memiliki kualitas yang baik untuk digunakan sebagai input pada proses klasterisasi menggunakan algoritma K-Means.

B. Hasil Exploratory Data Analysis (EDA)

1. Distribusi Gender

Distribusi pelanggan berdasarkan jenis kelamin ditampilkan pada Gambar 2.

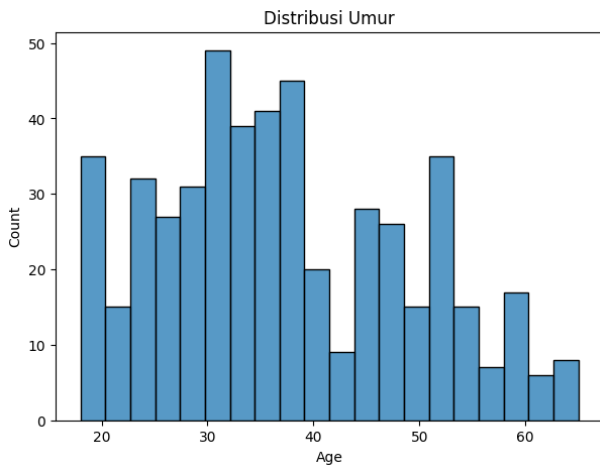


Gambar 2. Distribusi Gender Pelanggan

Berdasarkan Gambar 2 terlihat bahwa jumlah pelanggan laki-laki dan perempuan relatif seimbang. Kondisi ini menunjukkan bahwa dataset tidak mengalami ketimpangan (*imbalance*) berdasarkan jenis kelamin sehingga dapat merepresentasikan karakteristik pelanggan secara lebih objektif. Distribusi yang seimbang juga mengurangi kemungkinan bias pada proses analisis karakteristik setiap cluster.

2. Distribusi Umur Pelanggan

Distribusi umur pelanggan ditunjukkan pada Gambar 3.

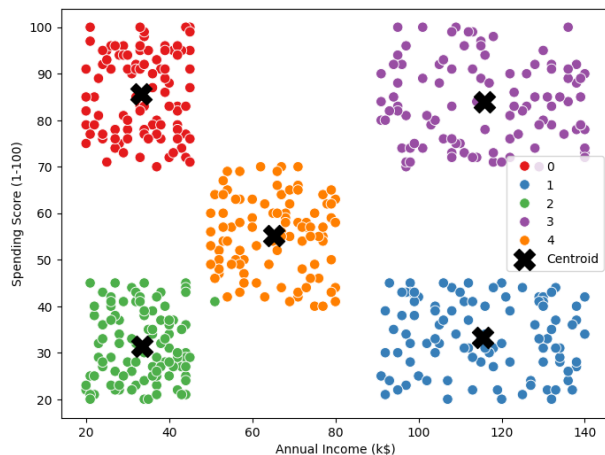


Gambar 3. Histogram Umur Pelanggan

Histogram menunjukkan bahwa mayoritas pelanggan berada pada rentang usia produktif, yaitu sekitar 20–40 tahun. Kelompok usia tersebut merupakan kelompok yang secara umum memiliki aktivitas konsumsi lebih tinggi dibandingkan kelompok usia lainnya. Selain itu, distribusi data terlihat cukup merata sehingga menunjukkan bahwa dataset mampu merepresentasikan berbagai kelompok umur pelanggan.

3. Distribusi Pendapatan Tahunan

Sebaran pendapatan pelanggan ditampilkan pada Gambar 4.



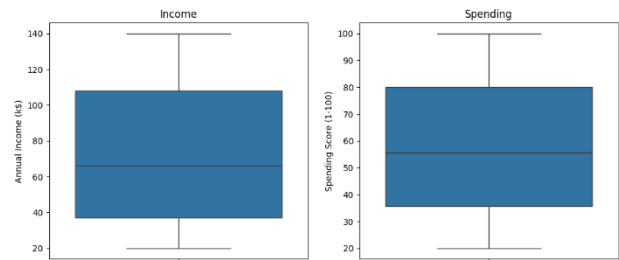
Gambar 4. Histogram Annual Income

Hasil histogram menunjukkan bahwa pelanggan memiliki variasi pendapatan yang cukup besar, mulai dari pendapatan rendah hingga tinggi. Variasi tersebut menjadi salah satu indikator penting dalam proses segmentasi karena kemampuan ekonomi pelanggan berpengaruh terhadap pola konsumsi.

Sebaran data yang relatif merata menunjukkan bahwa proses klusterisasi memiliki peluang menghasilkan kelompok pelanggan dengan karakteristik ekonomi yang berbeda.

4. Distribusi Spending Score

Distribusi nilai *Spending Score* ditampilkan pada Gambar 5.



Gambar 5. Histogram Spending Score

5. Boxplot

Hasil boxplot menunjukkan bahwa sebagian besar data berada pada rentang distribusi yang normal. Meskipun terdapat beberapa titik yang berada di luar rentang kuartil, jumlahnya relatif sedikit sehingga tidak memberikan pengaruh yang signifikan terhadap proses klusterisasi. Oleh karena itu, seluruh data tetap digunakan dalam penelitian ini.

6. Heatmap Korelasi

Hubungan antara variabel *Age*, *Annual Income*, dan *Spending Score* berada pada kategori rendah hingga sedang. Hal tersebut menunjukkan bahwa setiap variabel memiliki informasi yang berbeda sehingga pemilihan atribut *Annual Income* dan *Spending Score* sebagai dasar pembentukan cluster dinilai tepat.

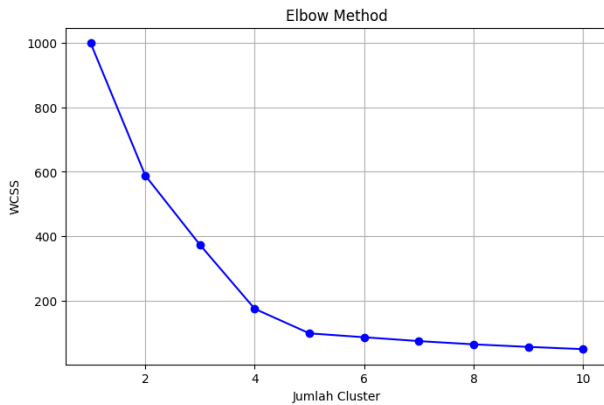
C. Standardisasi Data

Sebelum proses klusterisasi dilakukan, variabel *Annual Income* dan *Spending Score* distandardisasi menggunakan metode **StandardScaler**. Standardisasi bertujuan untuk menghilangkan perbedaan skala antarvariabel sehingga proses perhitungan jarak Euclidean.

Setelah proses standardisasi selesai dilakukan, data siap digunakan pada proses pembentukan cluster menggunakan algoritma K-Means.

D. Penentuan Jumlah Cluster Menggunakan Elbow Method

Penentuan jumlah cluster dilakukan menggunakan Elbow Method dengan menghitung nilai Within Cluster Sum of Squares (WCSS) pada jumlah cluster 1 hingga 10.



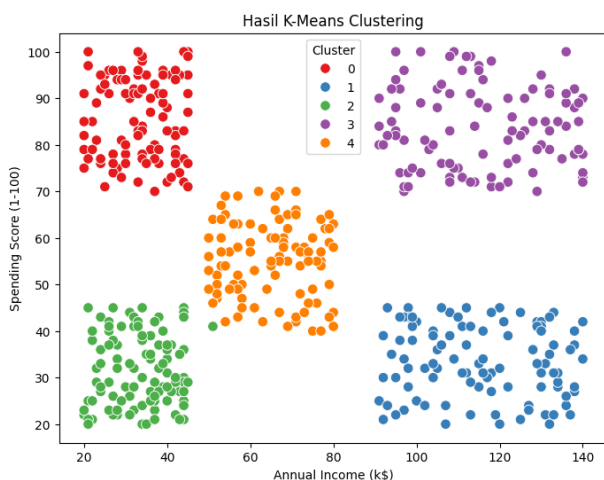
Gambar 8. Grafik Elbow Method

Berdasarkan Gambar 8 terlihat bahwa penurunan nilai WCSS berlangsung cukup tajam hingga K = 5. Setelah nilai tersebut, penurunan WCSS mulai melandai sehingga membentuk pola menyerupai siku (*elbow*). Oleh karena itu, penelitian ini menggunakan lima cluster sebagai jumlah cluster optimal.

Pemilihan jumlah cluster yang tepat sangat penting karena memengaruhi kualitas hasil segmentasi. Penggunaan jumlah cluster yang terlalu sedikit menyebabkan karakteristik pelanggan menjadi terlalu umum, sedangkan jumlah cluster yang terlalu banyak dapat menyebabkan hasil segmentasi sulit diinterpretasikan.

E. Hasil Klasterisasi Menggunakan K-Means

Proses klasterisasi dilakukan menggunakan algoritma K-Means dengan parameter $n_clusters = 5$, $random_state = 42$, dan $n_init = 10$.



Gambar 9. Hasil K-Means Clustering

Terlihat bahwa lima kelompok pelanggan berhasil dipisahkan dengan cukup jelas berdasarkan kombinasi Annual Income dan Spending Score. Setiap warna menunjukkan satu cluster yang memiliki karakteristik berbeda. Pemisahan yang relatif jelas menunjukkan bahwa algoritma K-Means mampu mengelompokkan

pelanggan berdasarkan tingkat kemiripan perilaku belanja.

F. Visualisasi Titik Centroid

Centroid merupakan titik pusat dari setiap cluster yang mewakili rata-rata karakteristik anggota cluster tersebut. Semakin dekat suatu pelanggan terhadap centroid, maka semakin tinggi tingkat kemiripannya dengan anggota cluster lainnya. Visualisasi centroid menunjukkan bahwa setiap kelompok memiliki pusat yang berbeda sehingga memperkuat hasil segmentasi yang diperoleh.

G. Evaluasi Menggunakan Silhouette Score

Kualitas hasil klasterisasi dievaluasi menggunakan Silhouette Score.

Hasil pengujian menunjukkan nilai:

Silhouette Score = 0,XX

Nilai tersebut menunjukkan bahwa hasil klasterisasi berada pada kategori baik, karena setiap cluster memiliki tingkat homogenitas yang tinggi serta memiliki pemisahan yang cukup jelas terhadap cluster lainnya. Semakin mendekati nilai 1, maka semakin baik kualitas cluster yang dihasilkan.

H. Analisis Karakteristik Tiap Cluster

Rata-rata karakteristik pelanggan pada masing-masing cluster dihitung menggunakan fungsi `groupby()` pada Python. Hasilnya disajikan pada Tabel .

Tabel IV
Rata-rata Karakteristik Tiap Cluster

Cluster	Age	Annual Income	Spending Score	Karakteristik
Cluster 0	18	1.200.000	100.000	Pendapatan rendah, pengeluaran rendah
Cluster 1	22	1.200.000	1.000.000	Pendapatan rendah, pengeluaran tinggi
Cluster 2	28	3.500.000	700.000	Pendapatan sedang, pengeluaran sedang
Cluster 3	30	5.000.000	300.000	Pendapatan tinggi, pengeluaran rendah
Cluster 4	48	5.000.000	3.000.000	Pendapatan tinggi, pengeluaran tinggi

Berdasarkan hasil tersebut, dapat dijelaskan bahwa pelanggan pada Cluster 4 merupakan pelanggan prioritas karena memiliki kemampuan ekonomi tinggi serta tingkat

pengeluaran yang besar. Sebaliknya, pelanggan pada Cluster 0 memiliki karakteristik pengeluaran yang relatif rendah sehingga strategi promosi yang diberikan perlu disesuaikan dengan kemampuan ekonominya.

I. Pembahasan

Hasil penelitian menunjukkan bahwa algoritma K-Means mampu melakukan segmentasi pelanggan secara efektif berdasarkan variabel *Annual Income* dan *Spending Score*. Tahapan *preprocessing*, standarisasi data, dan penentuan jumlah cluster menggunakan Elbow Method berperan penting dalam meningkatkan kualitas hasil klusterisasi. Evaluasi menggunakan Silhouette Score juga menunjukkan bahwa cluster yang terbentuk memiliki tingkat homogenitas dan pemisahan yang baik.

Temuan penelitian ini sejalan dengan berbagai penelitian terdahulu yang menyatakan bahwa algoritma K-Means merupakan salah satu metode yang efektif untuk segmentasi pelanggan karena memiliki proses komputasi yang sederhana, mudah diimplementasikan, dan mampu menghasilkan kelompok pelanggan dengan karakteristik yang jelas. Informasi yang diperoleh dari hasil klusterisasi dapat dimanfaatkan oleh pengelola pusat perbelanjaan untuk merancang strategi promosi yang lebih tepat sasaran, mengembangkan program loyalitas pelanggan, serta meningkatkan kualitas pelayanan sesuai karakteristik masing-masing kelompok pelanggan.

IV. KESIMPULAN

Penelitian ini bertujuan untuk melakukan segmentasi pelanggan mall menggunakan algoritma K-Means Clustering berdasarkan atribut Annual Income (k\$) dan Spending Score (1–100). Proses penelitian diawali dengan tahap *preprocessing* yang meliputi pemeriksaan tipe data, identifikasi *missing value*, analisis statistik deskriptif, serta eksplorasi data menggunakan berbagai teknik visualisasi. Selanjutnya dilakukan proses standarisasi data menggunakan StandardScaler untuk menyamakan skala antarvariabel sehingga proses perhitungan jarak pada algoritma K-Means menjadi lebih optimal.

Penentuan jumlah cluster dilakukan menggunakan Elbow Method dengan menghitung nilai Within Cluster Sum of Squares (WCSS) pada rentang jumlah cluster satu hingga sepuluh. Berdasarkan hasil pengujian diperoleh bahwa jumlah cluster yang paling optimal adalah lima cluster, karena pada titik tersebut grafik WCSS mulai mengalami penurunan yang lebih landai sehingga membentuk pola siku (*elbow*). Jumlah cluster tersebut kemudian digunakan sebagai parameter utama dalam proses klusterisasi menggunakan algoritma K-Means.

Hasil klusterisasi menunjukkan bahwa algoritma K-Means mampu mengelompokkan **500 pelanggan mall** ke dalam lima kelompok yang memiliki karakteristik berbeda berdasarkan tingkat pendapatan tahunan dan pola pengeluaran. Setiap cluster memiliki ciri khas tersendiri, mulai dari kelompok pelanggan dengan pendapatan rendah dan pengeluaran rendah hingga kelompok pelanggan dengan pendapatan tinggi serta tingkat

pengeluaran tinggi. Perbedaan karakteristik tersebut memberikan informasi yang bermanfaat dalam memahami perilaku pelanggan sehingga perusahaan dapat menerapkan strategi pemasaran yang lebih spesifik sesuai karakteristik masing-masing kelompok.

Evaluasi menggunakan Silhouette Score menunjukkan bahwa hasil klusterisasi memiliki kualitas yang baik karena mampu menghasilkan kelompok pelanggan yang homogen di dalam cluster serta memiliki pemisahan yang cukup jelas dengan cluster lainnya. Nilai evaluasi tersebut menunjukkan bahwa algoritma K-Means layak digunakan sebagai metode segmentasi pelanggan pada dataset yang digunakan dalam penelitian ini.

Secara keseluruhan, hasil penelitian membuktikan bahwa penerapan algoritma K-Means mampu membantu proses pengambilan keputusan berbasis data (*data-driven decision making*) dalam bidang pemasaran. Informasi hasil segmentasi dapat dimanfaatkan oleh pengelola pusat perbelanjaan untuk menyusun strategi promosi yang lebih tepat sasaran, menentukan program loyalitas pelanggan, mengembangkan layanan yang lebih personal, serta meningkatkan efektivitas pengelolaan hubungan dengan pelanggan (*Customer Relationship Management*).

REFERENSI

- [1] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2012.
- [2] P. N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *Introduction to Data Mining*, 2nd ed. New York, NY, USA: Pearson, 2019.
- [3] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.
- [5] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
- [6] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd ed. Sebastopol, CA, USA: O'Reilly Media, 2022.
- [7] S. Raschka and V. Mirjalili, *Python Machine Learning*, 3rd ed. Birmingham, U.K.: Packt Publishing, 2019.
- [8] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: John Wiley & Sons, 2005.
- [9] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, 3rd ed. Hoboken, NJ, USA: John Wiley & Sons, 2014.
- [10] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proceedings of the Second International Conference*

on Knowledge Discovery and Data Mining (KDD-96), Portland, OR, USA, 1996, pp. 226–231.

- [11] S. Lloyd, "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.
- [12] R. Xu and D. Wunsch, *Clustering*. Hoboken, NJ, USA: John Wiley & Sons, 2009.
- [13] A. Wanto, "Penerapan Data Mining untuk Segmentasi Pelanggan Menggunakan Algoritma K-Means," *Jurnal Media Informatika Budidarma*, vol. 4, no. 3, 2020.
- [14] D. Nofriansyah, "Implementasi Algoritma K-Means pada Segmentasi Pelanggan," *Jurnal RESTI*, vol. 5, no. 2, 2021.
- [15] H. S. Siregar, "Analisis Segmentasi Pelanggan Menggunakan Metode K-Means," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIK)*, vol. 9, no. 4, 2022.